QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Strategies for Manual Annotation of the Anopheles Genome

Kathryn Campbell

VectorBase

UNIVERSITY OF NOTRE DAME e! Ensembl IMBB Harvard EMBL

Anopheles gambiae
Manual Gene Annotation

Goal: Significantly improve the quality of gene annotations to reveal strategies for targeting *mosquito:parasite:human* interactions

- 278 Mb genome sequenced in 2002
- 14,707 Ensembl automated gene predictions
- 140,000 ESTs

The primary vector of malaria in Sub-Saharan Africa, the mosquito A. gambiae is a major contributor to the more than half-a-billion cases of malaria each year

Using Drosophila melanogaster as a paradigm, manual annotation of the Anopheles genome is expected to greatly improve the quality of annotations due to limitations of gene prediction algorithms.

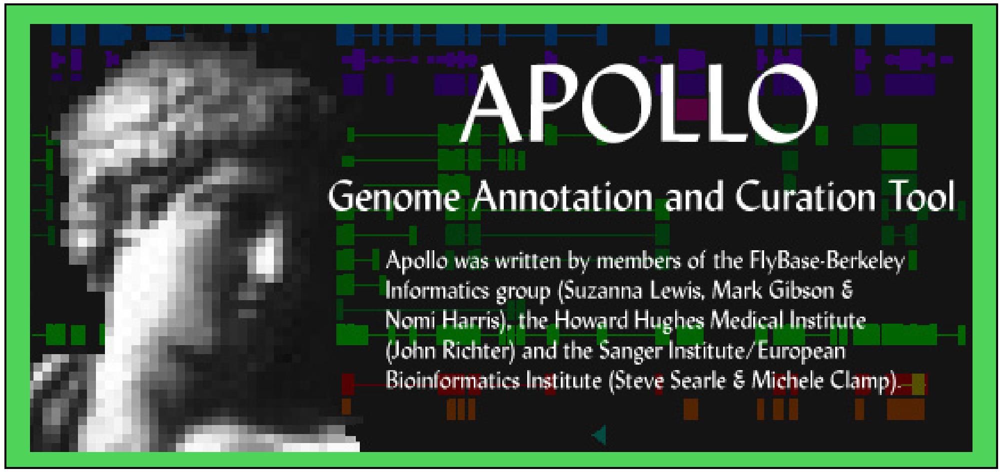
Strategy: Improve Annotations that represent computational challenges for automated gene builds

- **Alternative transcripts**
- **Overlapping genes**
- **Tandem homologous genes**
- **Dicistronic genes**
- **Nested genes**
- **Small ORFs**
- **Genes with repetitive sequence**

A first pass manual annotation will target regions and gene families of interest to the scientific community

Prioritizing Regions to be Annotated

- Genes involved in immune response, host preference, and insecticide resistance
- Multi-gene family members
- Tandem related genes
- 2 alternative assemblies of the same region due to divergent haplotypes in the PEST strain sequenced



- Apollo is the graphical interface that will be used for editing annotations on the Anopheles genomic sequence

- Apollo was used by FlyBase to manually annotate the entire 117Mb *Drosophila melanogaster* Euchromatin

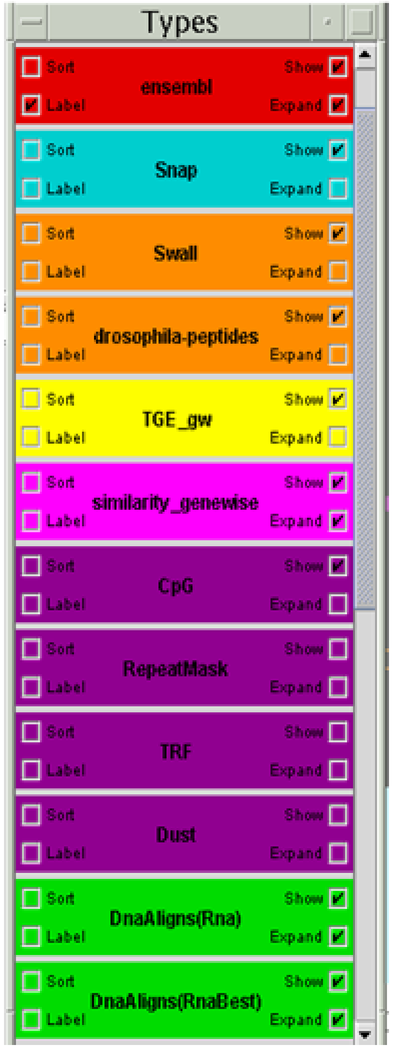
  (10 curators/7 months)

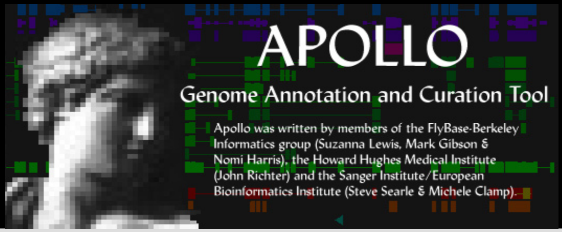
- Annotation editor for GMOD (Generic Model Organism Database)

- Developed by FlyBase-BDGP and Ensembl

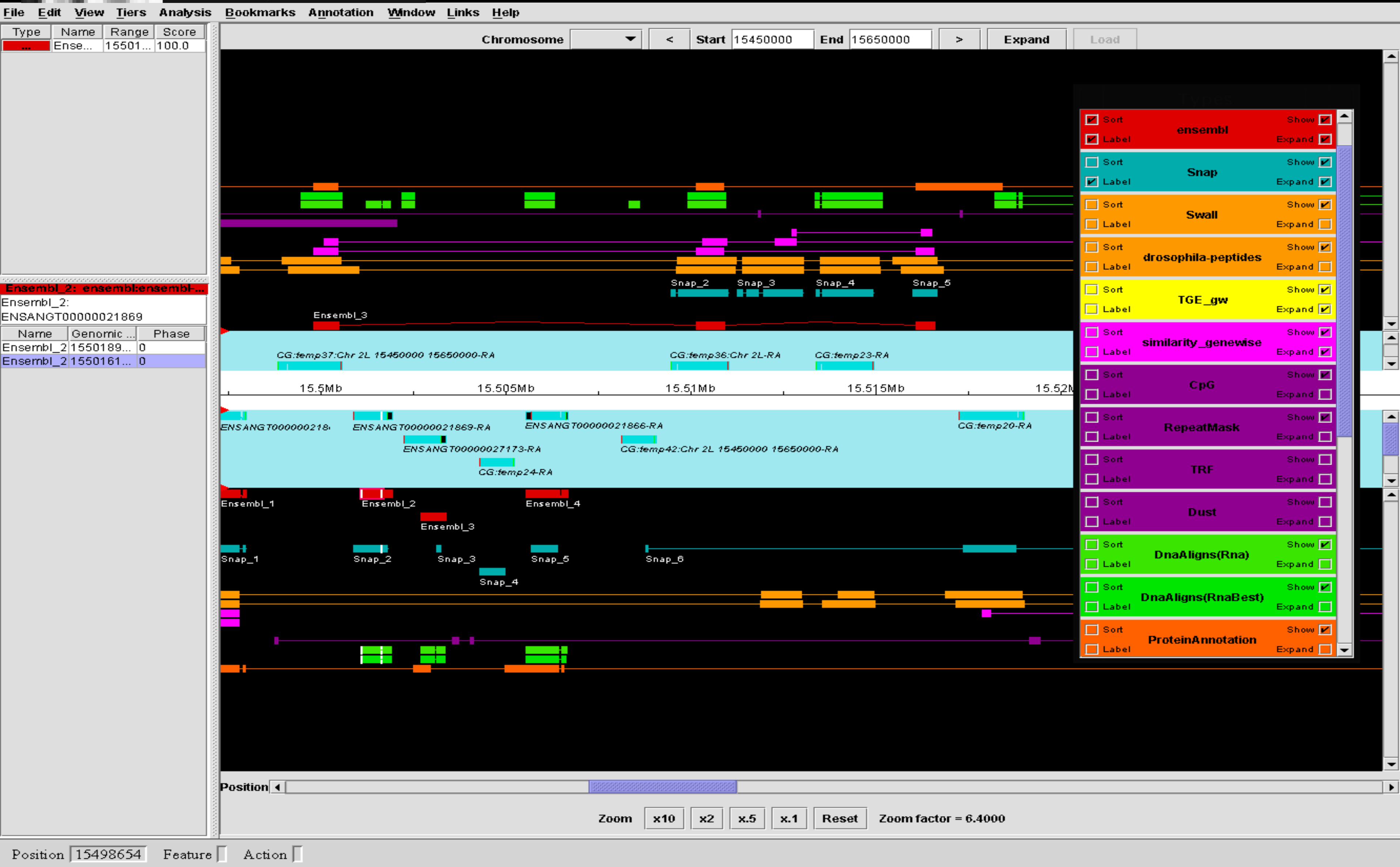
Data Types used for Manual Annotation

- Ensembl - Automated gene build based on synthesis of known proteins and cDNA/EST sequences
- SNAP- Gene finder trained on Anopheles
- TGE_gw - Anopheles/Drosophila protein aligned to the genome
- Genewise - Uniprot protein sequences aligned to the genome
- Genomewise - Ensembl EST transcript predictions
- SWALL - BLASTX similarity of SWISS-PROT/TrEMBL dataset
- Drosophila-Peptides - blast hit by Drosophila proteins.
- RNA_BEST - Anopheles EST/mRNA
- Protein features - Pfam domains, Prints family fingerprints, transmembrane regions, signal peptides
- RepeatMasker, TRF, SEG- identifies repeats/low complexity features

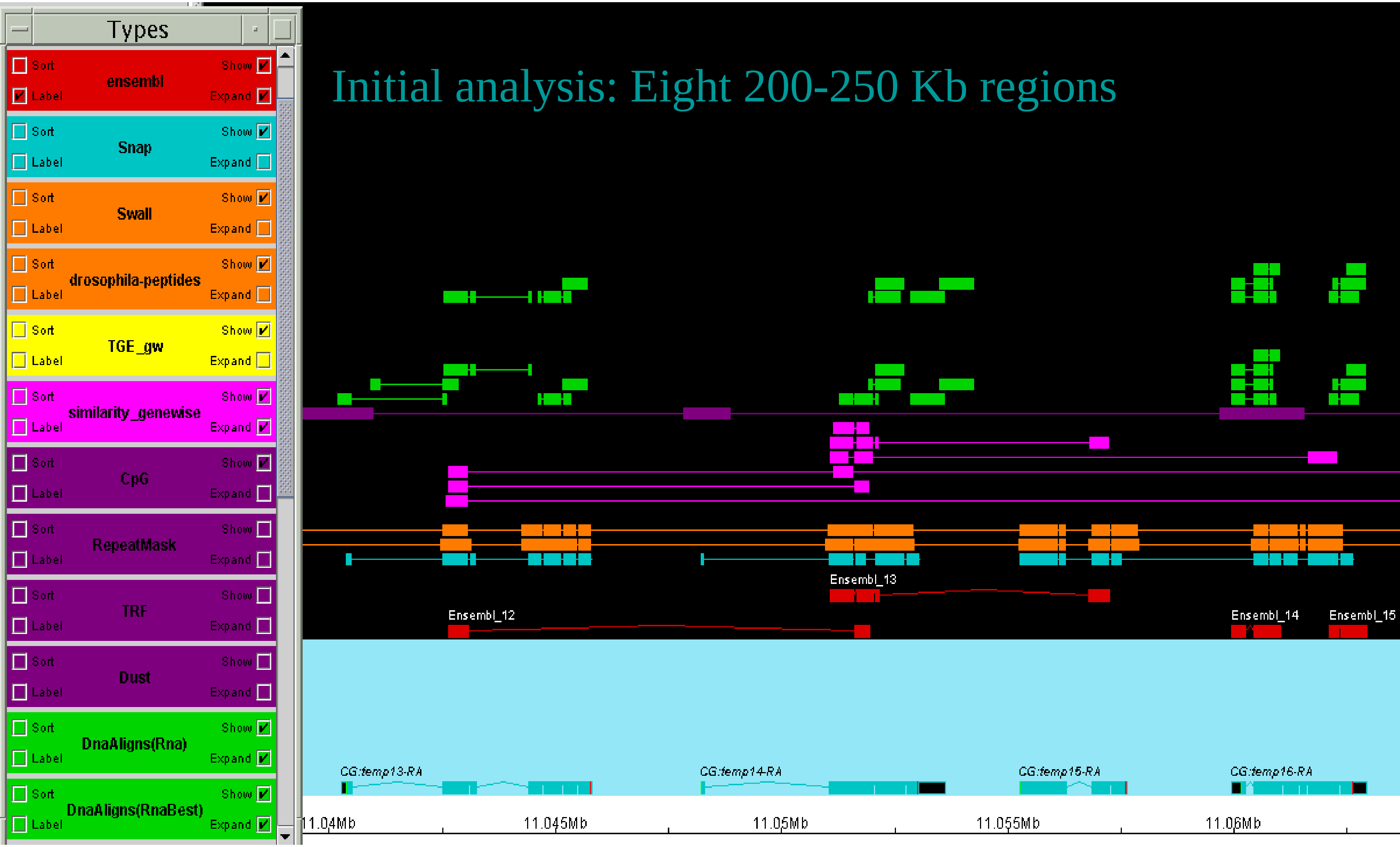




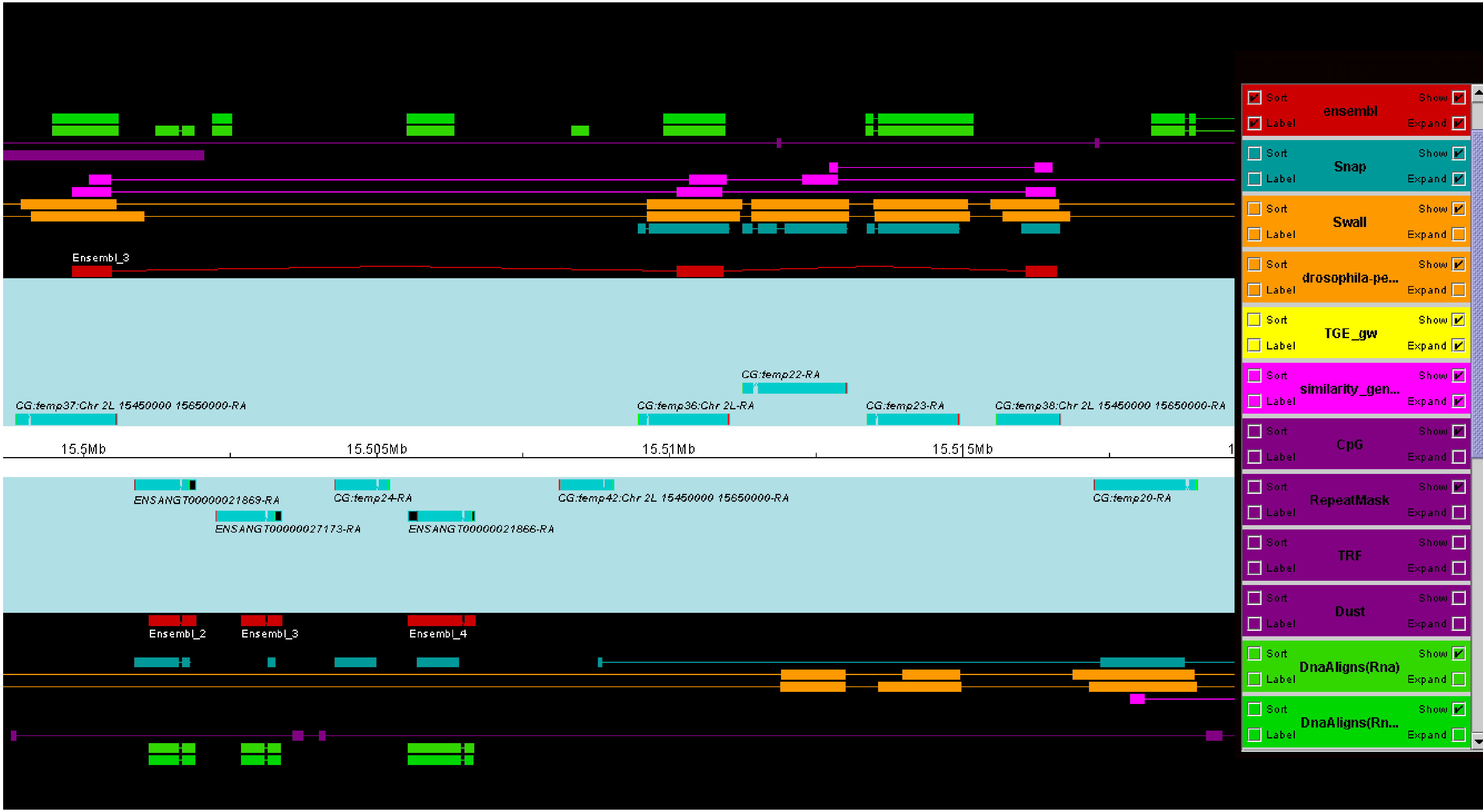
VectorBase curators will utilize the APOLLO graphic interface to view and synthesize data from multiple regions of the Anopheles genome



A curator is able to synthesize complex data sets resulting in significantly improved gene annotations

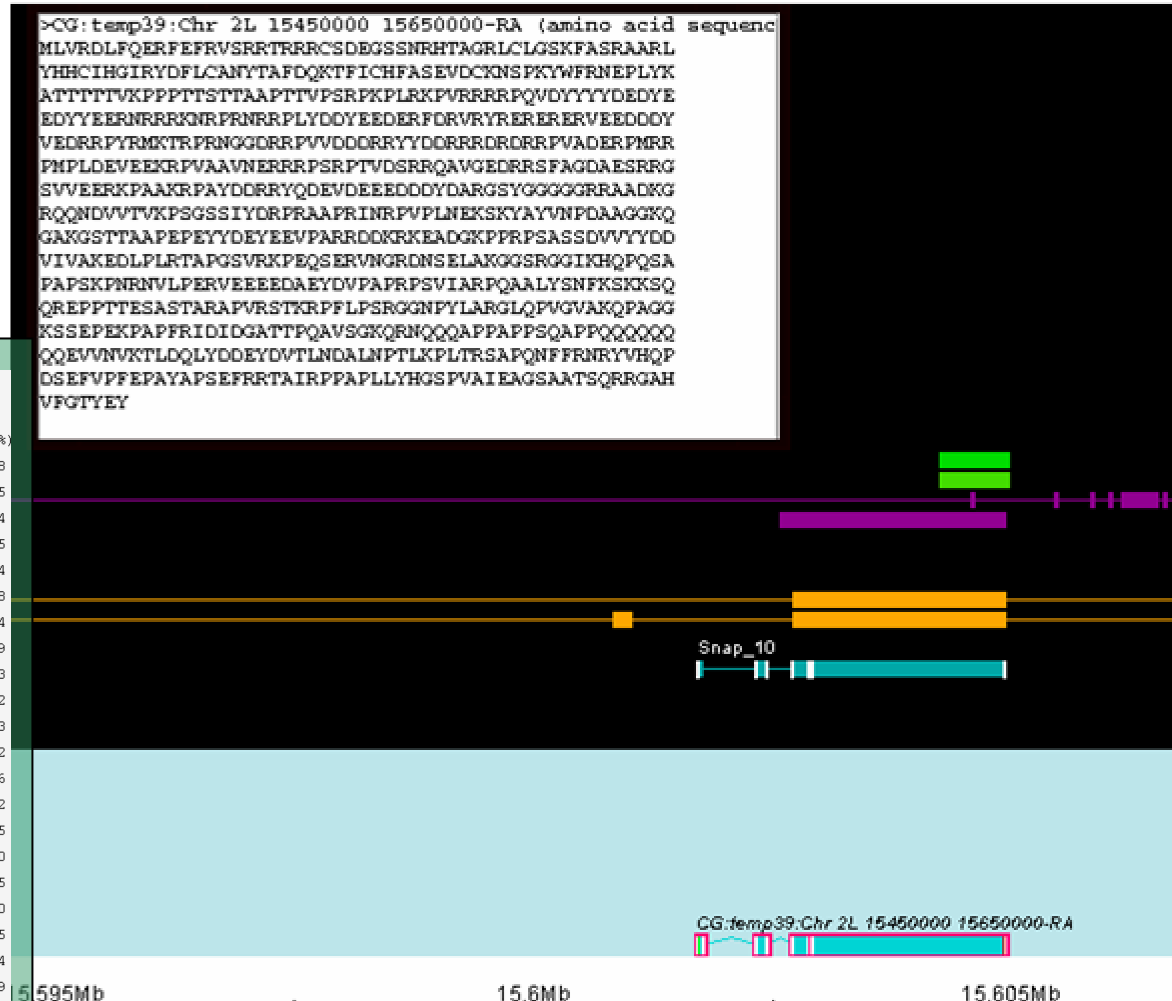


A Major Challenge for Automated Gene Annotation: Predicting Tandem Related Genes

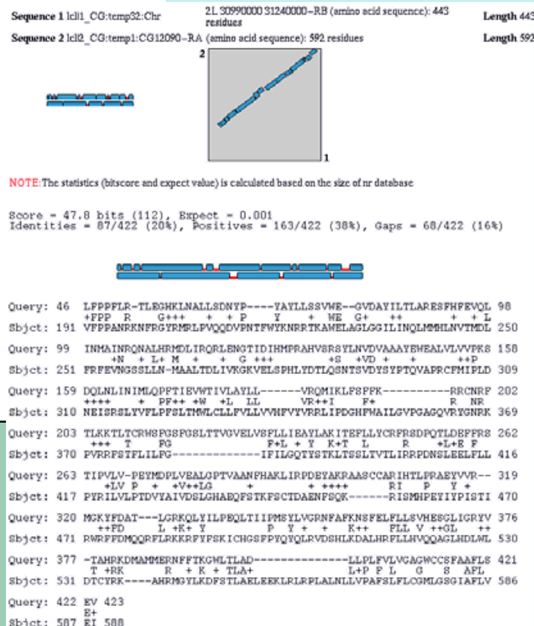


Low Complexity sequences may require manual annotation

```
>gi|28574954|ref|NP_648179.3| CG13676-PA [Drosophila melanogaster]
 gi|28380579|gb|AAF50480.2| CG13676-PA [Drosophila melanogaster]
          Length = 883

 Score =  166 bits (421), Expect = 2e-39
 Identities = 219/700 (31%), Positives = 304/700 (43%), Gaps = 144/700 (20%)

Query: 49  RLYHHCIHGIRYDFLCANYTAFDQKTFICHFASEVDCKNSPKYWFRNEPLYKATTTTTVK 108
           ++YHHC++GIR+DFLCAN+TAFDQ+TFICHFAS+VDC+ S KYW RN+ LY ATTTT+
Sbjct: 146 QIYHHCVYGIRHDFLCANFTAFDQRTFICHFASDVDCEGSQKYWNRNDELYMATTTTSTS 205

Query: 109 PPPTTSTTAAPTTVPSRPKPLRKPVR---RRRPQVDYYYYDED-YEEDYYEERNRRRKNR 164
              T +    P     RP+  ++P+R    RRP  DYYY D+D YE+DYYEER  RR
Sbjct: 206 TTTTPAPPTLPPAPRRRPQRPQRPLRRPYNRRPIDDYYYEDQDQYEDDYYEERPVRRPKP 265

Query: 165 PRNRRPLYDDYEEDERFDRVRYRERERERVEEDDDYVEDRRPYRMKTRPRNGGDRRPVVD 224
              +R    + E D+ +D  R   ++  R         +R  YR         D    +D
Sbjct: 266 RPRQRKPQVEPEYDDEYDDKRTEAKKPIR---------RNRNRYR---------DEEEALD 308

Query: 225 DDRRYYDDRRRDRDRRPVADERPMRRPMPLDEVEEKRPVAAVNERRRPSRPTVDSRRQAV 284
           +D                 DERP +RP         +P A    R+      +  R+
Sbjct: 309 EDY----------------DERPSKRP-------RGKPTAGTAGRK------ISPRKPGR 339

Query: 285 GEDRRSFAGDAE-SRRGSVVEERKPAAKRPAYDDRRYQDEVDEEEDDDYDARGSYGGGGG 343
            E+RRSF  D    RR S  E   P++     D+        E+  +  D   +
Sbjct: 340 VEERRSFNEDRPLGRRRSEKERTTPSSALDDLDEY-------EKPYNSADQVEAAEEQTP 392

Query: 344 RRAADKGRQQNDVVTVKPSGSSIYDRPRAAPRINRPVPLNEKSKYAYVNPDAAGGKQGAK 403
           ++     +   + +T K + +S+Y RPRA PRI RPVP+NEK K++Y         Q
Sbjct: 393 QKVKTTPKPLTEFITPKAAAASVYARPRAPPRIARPVPINEKKKFSYPVQKITATTQAPS 452

Query: 404 GSTTAA---PEPEYYDEYEEVPARRDDKRK----EADGKP-------------------- 436
            S       PE +  ++YE+ PAR   +R+     A+ KP
Sbjct: 453 NSAQEEEDYPEEQVEEDYEQPPARNTPRRRTPASRAEQKPTRTTLRKPVTDKKPVDEEYD 512

Query: 437 -----------PRPSASSDVVYYDDVIVAKEDLPLRTAPGSVRKPEQSERVNGRDNSELA 485
                       RP A    V   DV    E+     AP S   P +++R+  +  +
Sbjct: 513 EPAEPEPLRKRKRPLAPRSRVPAADVDFEDEEYEESPAPVSTAAPLRNQRLRAKTTTR-- 570

Query: 486 KGGSRGGIKHQPQSAPAPSKPNRNVLPERVEEEEDAEYDVPAPRPSVIARPQAALYSNFK 545
                     +P + P P KP      E     E  E + PAPR    +       + F
Sbjct: 571 ----------KPTAPPRPRKPIVEDEEELEHPVEAVEDEAPAPRSRANSLSGRNAGAGFS 620

Query: 546 SKKSQQREPPTTESASTARAPVRSTKRPFLPSRGGNPYLARGLQPVGVAKQPAGGKSSEP 605
            + S                 VR  KRPFLPSRGG+PYL RGLQPVGVA +P     S P
Sbjct: 621 QRPSS----------------VRVVKRPFLPSRGGSPYLPRGLQPVGVALKPLPTTDSTP 664

Query: 606 -------------EKPAPF---RIDIDGATTPQAVSGKQRNQQQAPPAPPSQAPPQQQQQ 649
                        E  AP      D        A       +   PP     A P++ +
Sbjct: 665 IDMGSTISGVRLLEHGAPLLRGNPDSGEDDEEDAPPSPSPPRTTLPPRSALPATPKRVEP 724

Query: 650 QQQEVVNVKTLDQLYDDEYDVTLNDALNPTLKPLTRSAPQ 689
           Q+ ++     LD+LY+++YDVTLNDAL+PTLKPL+   P
```



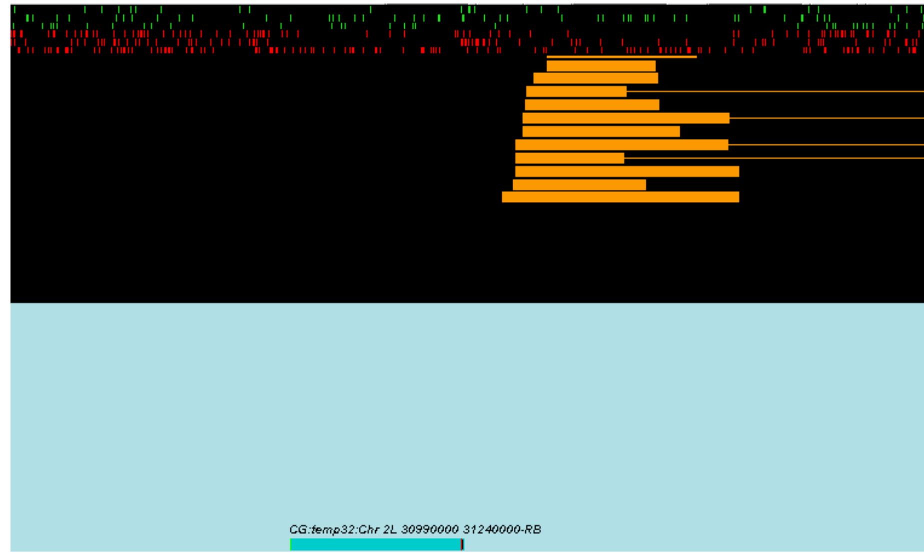
Finding novel genes requires manual annotation



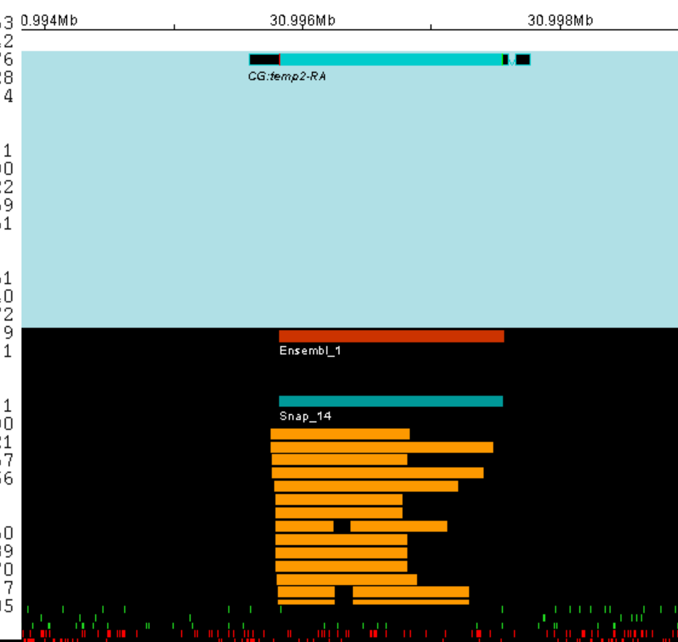
```
Sequence 1 lcl|1_CG:temp32:Chr   2L 30990000 31240000-RB (amino acid sequence): 443 residues   Length 443
Sequence 2 lcl|2_CG:temp1:CG12090-RA (amino acid sequence): 592 residues                     Length 592
```

NOTE: The statistics (bitscore and expect value) is calculated based on the size of nr database

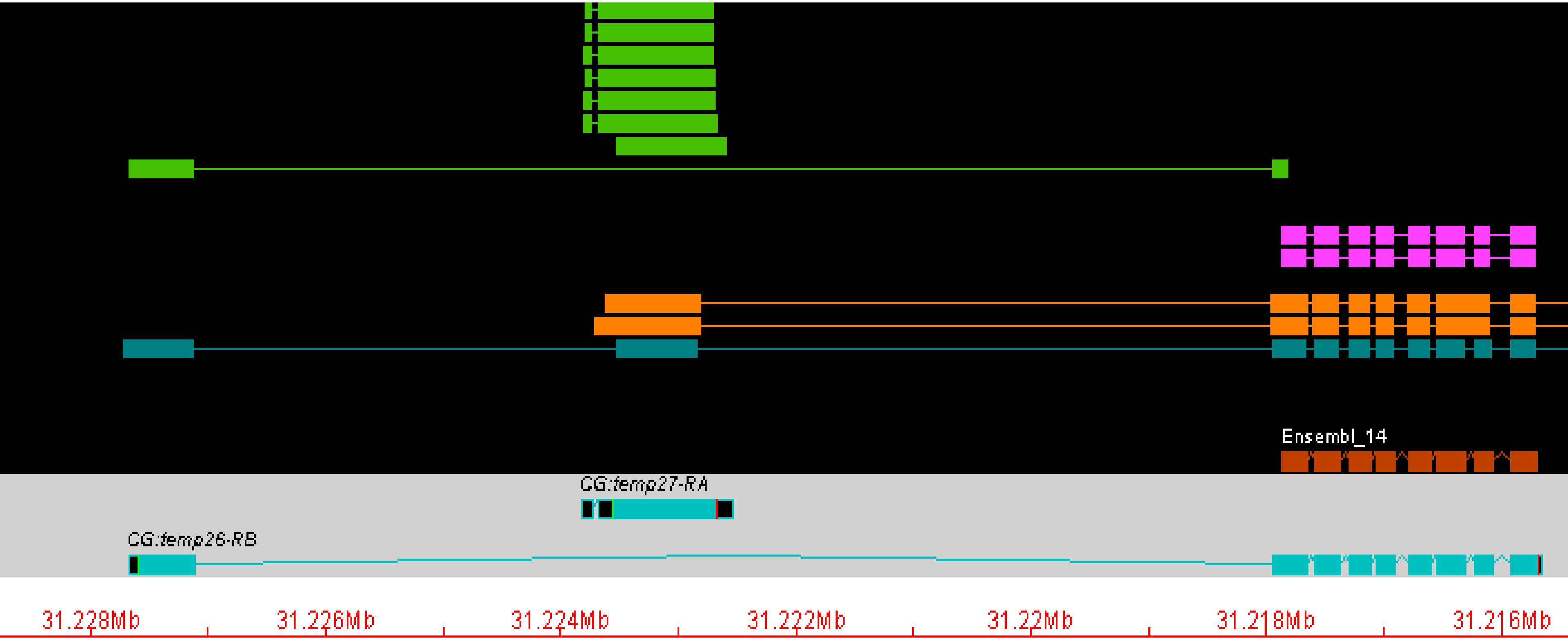
```
Score = 47.8 bits (112), Expect = 0.001
Identities = 87/422 (20%), Positives = 163/422 (38%), Gaps = 68/422 (16%)

Query: 46  LFPPFLR-TLEGHKLNALLSDNYP----YAYLLSSVWE--GVDAYILTLARESFHFEVQL 98
            +FPP  R    G+++    +  + P    Y     +  WE  G+    ++          +  + L
Sbjct: 191 VFPPANRKNFRGYRMRLPVQQDVPNTFWYKNRRTKAWELAGLGGILINQLMMHLNVTMDL 250

Query: 99  INMAINRQNALHRMDLIRQRLENGTIDIHMPRAHVSRSYLNVDVAAAYEWEALVLVVPKS 158
                +N  + L+ M  +    +  G +++           +S  +VD +   +        ++P
Sbjct: 251 FRFEVNGSSLLN-MAALTDLIVKGKVELSPHLYDTLQSNTSVDYSYPTQVAPRCFMIPLD 309

Query: 159 DQLNLINIMLQPFTIEVWTIVLAYLL------VRQMIKLFSPFK----------RRCNRF 202
            ++++    +  PF++ +W  +L  LL       VR++I    F+          R  NR
Sbjct: 310 NEISRSLYVFLPFSLTMWLCLLFVLLVVMFVYVRRLIPDGHFWAILGVPGAGQVRYGNRK 369

Query: 203 TLKKTLTCRWSFGSFGSLTTVGVELVSFLLIEAYLAKITEFLLYCRFRSDPQTLDEFFRS 262
             +++  T    FG                 F+L + Y  K+T  L     R    +L+E F
Sbjct: 370 PVRRFSTFLILFG--------------IFILGQTYSTKLTSSLTVTLIRRPDNSLEELFLL 416

Query: 263 TIPVLV-PEYMDPLVEALGPTVAANFHAKLIRPDEYAKRAASCCARIHTLPRAEYVVR-- 319
              +LV P  +  +V++LG     +        + ++++      RI   P   Y +
Sbjct: 417 PYRILVLPTDVYAIVDSLGHAEQFSTKFSCTDAENFSQK------RISMHPEYIYPISTI 470

Query: 320 MGKYFDAT---LGRKQLYILPEQLTIIPMSYLVGRNFAFKNSFELFLLSVHESGLIGRYV 376
             ++FD     L +K+ Y          P  Y +  +   K++   FLL V ++GL   ++
Sbjct: 471 RWRFFDMQQRFLRKKRFYFSKICHGSFPYQYQLRVDSHLKDALHRFLLHVQQAGLHDLWL 530

Query: 377 -TAHRKDMAMMERNFFTKGWLTLAD-------------LLPLFVLVGAGWCCSFAAFLS 421
             T +RK    R + K + TLA+                 L+P F L   G   S  AFL
Sbjct: 531 DTCYRK----AHRMGYLKDFSTLAELEEKLRLRPLALNLLVPAFSLFLCGMLGSGIAFLV 586

Query: 422 EV 423
            E+
Sbjct: 587 EI 588
```




```
ENSANGP00000029595                VATRDMVVLDHA--PKLHLTSRDYFNLVVPKPVKLNLIDAMVQPFNGTVW 263
ENSANGP00000025936                VATRDMVGLDHT--PKLHLTSRDYFNLVVPKPVKLNLIDAMVQPFNGTVW 342
CG_temp32_Chr2L30990000312400000  HMPRAHVSRSYLNVDVAAAYEWEALVLVVPKSDQLNLINIMLQPFTIEVW 176
ENSANGP00000028419                MMTRRHVNVGIL--PVVYIPDITYYCLVAPRTTQIDLTQSLLRPFSGTVW 228
ENSANGP00000026593                DLTFAYTSFRAHYMHDVSLKERGGYCVLCPFHTERDFLRHLLKPFSFGIW 214
                                   .   .           .      :: *    :  ::    :::**.  :*

ENSANGP00000029595                LMIGALLGVRVFAGYLQDLFGMLDRSGTLRKWLRSLQALHFP--CPQWVQ 311
ENSANGP00000025936                LMIGVLLGVRVFAGYLQDLFGLLDRSGKVRKWWRSLQAPHCP--CTQWVQ 390
CG_temp32_Chr2L30990000312400000  TIVLAYLLVRQMI----KLFSFFKRRCNRFTLKKTLTCRWSFGSFGSLTT 222
ENSANGP00000028419                WFIVVCTVL---------ISAFDELSKQHTRLGRLAQQLFARQPIASFYR 269
ENSANGP00000026593                AVLGALLVGCRLLG---HLFPALFERNLLEQIFFTAGASHRQPFPTRIVS 261
                                   .:  .           :

ENSANGP00000029595                LAADVLTFLLIEAYLAQVTSLLLTLRFIEGPRDLNEFIASNIRIVEPYES 361
ENSANGP00000025936                LAADVLTFLLIEAYLAQVTSLLLTLRFIEGPRDLNEFIASNMRIVEPYES 440
CG_temp32_Chr2L30990000312400000  VGVELVSFLLIEAYLAKITEFLLYCRFRSDPQTLDEFFRSTIPVLVPEYM 272
ENSANGP00000028419                ICLAVISFVLIESYLATVTSFFLAYRFVPDAKTLEEFFATGIPIRLPEGM 319
ENSANGP00000026593                FSAAVLIFFLSEAYNAKIVSLMSDSKYFDRPESVRELIESDLKVAIPGVR 311
                                   .    :: *.*  *:*  *  :..::    ::    ..  : *:: :  :  :   *

ENSANGP00000029595                TLSLVQVETGQRALLKTRFVKRTAEELARPNQDDAFVELRSRVAFQSYGT 411
ENSANGP00000025936                TLSLVQVETGQRALLKTRFVKRTAEELARPNQDDAFVELRSRVAFQSYGT 490
CG_temp32_Chr2L30990000312400000  DPLVEALGPTVAANFHAKLIR-PDEYAKRAASCCARIHTLPRAEYVVRMG 321
ENSANGP00000028419                VRFLRNLEPRLKDRIMARGVG--ASVCSPYSTQCAHLESFAMASYQISEN 367
ENSANGP00000026593                ASLLAESLPGKLVNRRR-----AEALYRERGPLLFNEYCTVMYCYLAYLQ 356
                                     :      .                                      :

ENSANGP00000029595                EPFDPVTGRRNFYLLDEPLADVR-FQYSFAKETAFMGVAQDYMLRCEENG 460
ENSANGP00000025936                EPFDPVTGRRNFYLLDEPLADVR-FQYSFAKETAFMGVAQDYMLRCEENG 539
CG_temp32_Chr2L30990000312400000  KYFDATLGRKQLYILPEQLTIIP-MSYLVGRNFAFKNSFELFLLSVHESG 370
ENSANGP00000028419                VGIDPVSGRKRSYIVPEMVTRYSYLSYAFARGSPLTDLVAMYLRWMYEAG 417
ENSANGP00000026593                TSVGRNLHGFQQYVLPDMVSEQL-RTLQLATHSPFYDTFAEYYERYFQSG 405
                                    ..       . *::  :  ::           ..  .:  .        :       :  *
```

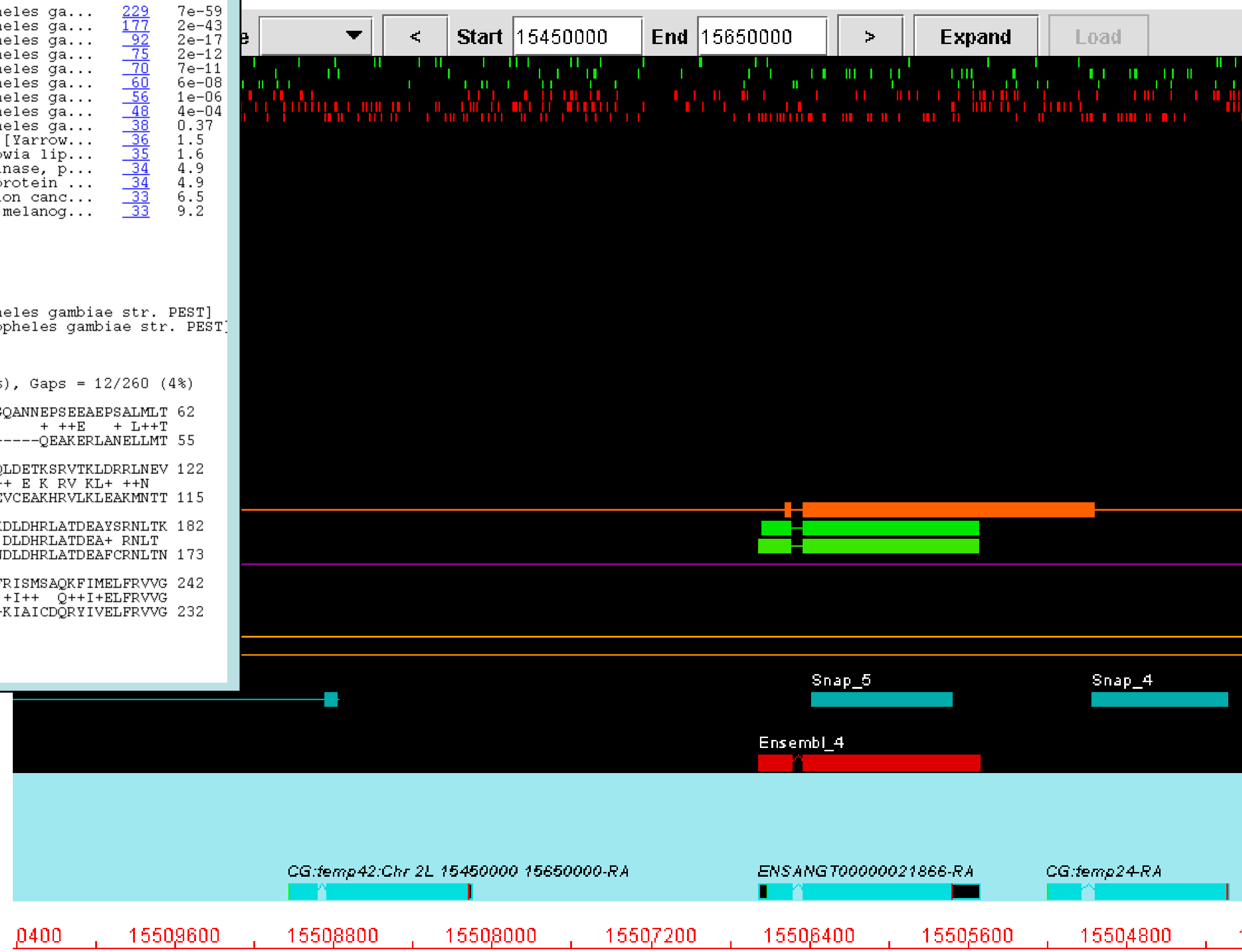


Nested genes pose a challenge for gene prediction algorithms



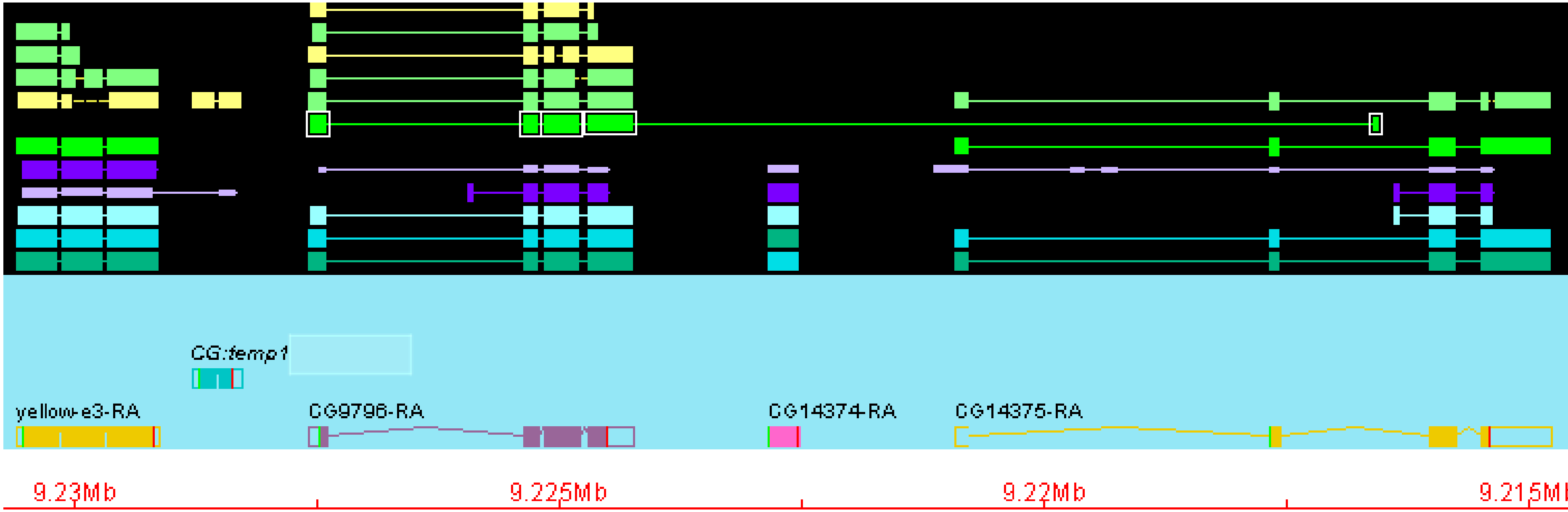
Putative Anopheles specific gene identified by pattern recognition and visual search for ORFs

```
                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gi|55238289|gb|EAA11136.3|  ENSANGP00000021866 [Anopheles ga...   229   7e-59
gi|55238287|gb|EAA11135.2|  ENSANGP00000021869 [Anopheles ga...   177   2e-43
gi|55238288|gb|EAL39836.1|  ENSANGP00000027562 [Anopheles ga...    92   2e-17
gi|55239049|gb|EAL40035.1|  ENSANGP00000025919 [Anopheles ga...    75   2e-12
gi|55238124|gb|EAL39790.1|  ENSANGP00000025599 [Anopheles ga...    70   7e-11
gi|55240053|gb|EAL40329.1|  ENSANGP00000029231 [Anopheles ga...    60   6e-08
gi|55244958|gb|EAL41642.1|  ENSANGP00000026714 [Anopheles ga...    56   1e-06
gi|55241665|gb|EAL40761.1|  ENSANGP00000028587 [Anopheles ga...    48   4e-04
gi|55244020|gb|EAL41360.1|  ENSANGP00000029329 [Anopheles ga...    38   0.37
gi|49646778|emb|CAG83163.1|  unnamed protein product [Yarrow...    36   1.5
gi|297855|emb|CAA42211.1|  fatty-acid synthase [Yarrowia lip...    35   1.6
gi|56468200|gb|EAL46075.1|  phosphatidylinositol 3-kinase, p...    34   4.9
gi|56459878|ref|YP_155159.1|  Thiamine biosynthesis protein ...    34   4.9
gi|34067911|gb|AAQ56721.1|  serologically defined colon canc...    33   6.5
gi|24662899|ref|NP_648506.1|  CG14131-PA [Drosophila melanog...    33   9.2


                               Alignments


>gi|55238289|gb|EAA11136.3| ENSANGP00000021866 [Anopheles gambiae str. PEST]
 gi|58387327|ref|XP_315479.2| ENSANGP00000021866 [Anopheles gambiae str. PEST]
          Length = 289

 Score =  229 bits (584), Expect = 7e-59
 Identities = 132/260 (50%), Positives = 180/260 (69%), Gaps = 12/260 (4%)

Query: 3    SSRDKRSGSIYKFRDALIKKIEKKFEEEERLLRLQRKLVVLGQANNEPSEEAEPSALMLT 62
            S  DK +G IY+FRD L  K   +FE E++L+RLQR+          + ++E   + L++T
Sbjct: 5    SQHDKDTGKIYRFRDKLESKFNAEFEAEQKLIRLQRE---------QEAKERLANELLMT 55

Query: 63   CDDEDPFDADPSNATTADVVKVLRTLSVQIEQLSCKQDTFQHQLDETKSRVTKLDRRLNEV 122
            C++EDP  +P +AT ADV+K+L+T+S +I QL  KQDTFQ ++ E K RV KL+ ++N
Sbjct: 56   CEEEDPTINPDSATIADVMKMLKTVSRKIGQLCGKQDTFQKEVCEAKHRVLKLEAKMNTT 115

Query: 123  LFMAEAVKDELMFDDFQKATEQIVTPVPWFEFQAVSNETELKDLDHRLATDEAYSRNLTK 182
            L M E VKDE++    +    EQ    P P FEF+AVSNE E+ DLDHRLATDEA+ RNLT
Sbjct: 116  LSMTEKVKDEMLARYW--TPEQDDPPEPGFEFRAVSNEQEMNDLDHRLATDEAFCRNLTN 173

Query: 183  WLNAKILHPDPNHRLHQAMEAVFKREFLPSCSWKGRAKPGQFRISMSAQKFIMELFRVVG 242
            WL AKI     P  R+   A++ VF+REFL   CSWKGR+K G   +I++   Q++I+ELFRVVG
Sbjct: 174  WLKAKISLQKPFRRMLHALDLVFEREFLMLCSWKGRSKSGP-KIAICDQRYIVELFRVVG 232

Query: 243  SNRFITISDEFVAKFFIKKL 262
            SNR  TI+D+ VAKFF++KL
Sbjct: 233  SNRCTTITDKIVAKFFVRKL 252
```



CGtemp42 and CGtemp24 are both homologues of ENSANG_21866

Manual annotation of Anopheles leads to annotation of new Drosophila genes



CLUSTAL W alignment with DPM3 proteins

New Dmel gene
New Dpse gene
New Agam gene

```
CLUSTAL W (1.82) multiple sequence alignment


DPM3-Bos                           MTKLAQWLWALALLGSTWAALTMGALGLELPSSCREVLWPLPAYLLVSAG 50
DPM3-Mus                           MTKLTQWLWGLALLGSAWAALTMGALGLELPFPCREVLWPLPAYLLVSAG 50
DPM3-Homo                          MTKLAQWLWGLAILGSTWVALTTGALGLELPLSCQEVLWPLPAYLLVSAG 50
DPM3-Xenopus                       MTKLAEWLLALSVLGAAWVTLNFNLLGLDLPPPLQQLLWPLPVYLLVVFG 50
CG_temp1-Dmel                      MTNLQRWLFYASLFAIPYLSVVLGTVQTPLTTKYFLHIQLLPLLLLVIFG 50
Dpse_Contig1107_Contig3589         MTNLQRWLFYATLFAVPYLSVVLGTVQTQFTSKYLLHIQLLPLLLLVLFG 50
CG_temp4Anopheles                  MTKLFEWFMAAACFFSVYFAIVLRQVKHPLLDEYMLEIQLSPLFLVLLFG 50
                                   **:* .*:   : :   : ::    :   :       :   *  *::  *

DPM3-Bos                           CYALGTVGYRVATFHDCEDAARELQSQIQEARADLTRRGLRF-- 92
DPM3-Mus                           CYALGTVGYRVATFHDCEDAARELQSQIVEARADLARRGLRF-- 92
DPM3-Homo                          CYALGTVGYRVATFHDCEDAARELQSQIQEARADLARRGLRF-- 92
DPM3-Xenopus                       CYSLATIGYRVATFNDCEDAARELQQQISEAKRDLALKGLKF-- 92
CG_temp1-Dmel                      IYSVWTVLYRTLTFNDCPEAAKELQDEIQEARKDLIAKGFRFRD 94
Dpse_Contig1107_Contig3589         LYSVWTVLYRTFTFNDCPEAAKELQAEILEARKDLIAKGFRFRD 94
CG_temp4Anopheles                  IFSATVVLYRTFTFNNCEEAAKELMEQIKEAKADLRSKGLVLSD 94
                                    ::  .: **. **::* :**:**  :* **: **  :*: :
```

Results of *Anopheles gambiae* Manual Annotation

Initial Analysis of 1.9 Mb and 161 Ensembl annotations

2L	15450000 - 15650000
2L	17750000 - 18000000
2L	30990000 - 31240000
2R	14450000 - 14650000
2R	32210000 - 32460000
2R	59310000 - 59510000
3L	10900000 - 11120000
3L	11650000 - 11900000

- 49 Annotations required either additional coding exons, 5’ or 3’ sequence based on blast or EST or alternative transcripts
- 28 New genes annotated based on Gene prediction, ESTs and BLASTX
- 7 Gene merges - 14 Ensembl annotations merge into 7
- 10 Gene splits - 10 Ensembl annotations split into 22 new annotations
- 7 Ensembl annotations designated for deletion
- 25 Ensembl annotations required only minor adjustment to 5’ and/or 3’ sequence
- 56 Ensembl annotations required no change.

Results of *Anopheles gambiae* Manual Annotation

Initial Analysis of 1.9 Mb and 161 Ensembl annotations

2L	15450000 - 15650000
2L	17750000 - 18000000
2L	30990000 - 31240000
2R	14450000 - 14650000
2R	32210000 - 32460000
2R	59310000 - 59510000
3L	10900000 - 11120000
3L	11650000 - 11900000

- 50% of the genes predicted by the Ensembl Automatic Gene Building System required little or no change
- 30% of the genes required moderate changes to the gene structure
- 15% of automatic gene annotations were either split or merged (primarily based on BLAST evidence)
- The 28 new genes annotated comprised tandem related genes, nested genes, genes with low complexity sequence.

Limitations, challenges and the future of manual gene annotation

- **While manual annotation provides the highest quality gene annotations, it is both labor intensive and expensive.**

- **The data illustrate the challenges for automated gene builds. Annotators will work together with Ensembl to rigorously analyze discrepancies between automated and manual Anopheles gene structures in an effort to improve gene prediction algorithms.**

- **These efforts will inform future automated annotation pipelines of not only Anopheles but also Aedes.**

- **Not unexpectedly, the initial pilot manual annotation Anopheles has led to corrections of several Drosophila melanogaster gene structures.**

VectorBase

Ewan Birney
Martin Hammond
Vivek Iyer
Bill Gelbart
Frank Collins
Christos Louis
Fotis Kafatos

UNIVERSITY OF NOTRE DAME   e! Ensembl   IMBB   Harvard   EMBL